1 Domain Background

The proposed project centers around the art of counterfeiting banknotes. Know colloquially as *worlds second oldest profession*, this art dates back to the invention of money itself. From counterfeiters reproducing gold coins in Roman times using base metals, to the use of Mulberry trees to imitate banknotes 13th century China[1], counterfeiters have been playing a fiendish cat-and-mouse game with governments that makes for a fascinating history. Counterfeiting techniques were even used as a form of warfare in the U.S. Revolutionary war of independence by the British in order to devaluate the newly created dollar [1].

The problem of counterfeiting is not just one of historical importance however. In fact it is estimated that counterfeiting produces an annual increase in inflation resulting in a 250 billion dollar loss annually for the us economy [2].

In more modern times, the techniques used have grown very subtle as anti-counterfeiting measures have entered into the 21st century. The main way to detect counterfeit money has however remained unchanged: the note is ran through a subsequent series of tests to determine any possible foul play.

2 Problem Statement

The problem this proposal wishes to consider is to determine whether it is possible to build an algorithm that detects foul play simply by scanning a bill, and extracting any anomalies in the resulting image.

More specifically, we will encode the image of a bill into a set of 5 data points which correspond to statistical features associated with the image and train a learning algorithm capable of classifying real bills from counterfeit ones.

3 dataset

The dataset in question was obtained as from the UCI machine learning repository[3]. It serves as the basis for a regression analysis performed by Gillich and Lowesh in [GLon]. In loc. sit. the authors first scanned a set of 1372 banknotes of which 610 were forged into 400× 400 pixel images. To each image, a *Wavelet transform* was applied. This is a mathematical tools which finds its origins in the theory of Fourier transforms and allows one to encode the image data very efficiently. For our purposes it will be sufficient to note that it compresses an image by extracting coefficients

[1] en.wikipedia.org/wiki/Counterfeit

[2] http://www.ipwatchdog.com/2010/08/30/counterfeiting-costs-us-businesses/id=12336/

[3] https://archive.ics.uci.edu/ml/datasets/banknote+authentication

which follow an underlying probability distribution. The data set in question consists of a description of 4 statistical properties of this distribution:

1. the variance
2. the sample skewness
3. the kurtosis
4. the Shannon entropy

Together with a binary target variable indicating whether the note is real or counterfeit

4 Solution Statement

Since the four feature variable all describe an aspect of the the shape of the distribution that encodes the compressed image, we expect an appropriate linear combination of this features to be indicative of whether a banknote is counterfeit. Our first proposed solution will thus take form of a logistic classifier.This classifier its speed and simplicity, is considered an industry workhorse for binary classification problems [4]. Although this particular problem seems to be a tad singular in nature, many related problems have been treated through the use of logistic regression: mortality risk assessment based off the descriptive statistics of cardiac data[5] or the detection of loan defaults based off demographics [6] to name but two.
We will also fit an ID3 classifier to account for the possibility of a nonlinear relation. Finally, we will cluster the data using a support vector machine.

5 Evaluation Metrics

To evaluate the performance of the model, we intend to use the tools available to us through scikit-learn. More precisely, the will be measured be analyzing precision, recall and the f1 score of the classifier.

6 Benchmark Model

As a benchmark model, we will train and test our data by running k-nearest neighbor out of the box. After finetuning the various classifiers proposed in §4,wWe will relate

[4] https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-algorithm-choice

[5] http://ieeexplore.ieee.org/document/6420438/

[6] https://smartdrill.com/pdf/Credit%20Risk%20Analysis.pdf

the evaluation metrics against the benchmark. This will give and objective way to compare the improvement made by each classifier as it gets calibrated properly.

7 Project Design

The workflow of the project will be subdivided as follows:

7.1 Data Preprocessing

As a preliminary step, we will provide an overview of the data set through descriptive statistics (min/max/quartiles etc) and exhibit a scatterplot of correlations. This information will allow us to preprocess the data accordingly by applying any necessary feature scaling, removing correlated variables if any and removing outliers if any.

7.2 Fitting the Classifiers

As a next step, we will build a logistic classifier.The data will be randomized and split into training and testing sets (75%-25%).
Subsequently, we plan on fitting the various classifiers (logistic, ID3) to the data and finetuning. To this end, we will analyze how the data gets either over- or underfitted. Based of these conclusions, we may choose to optimize the regularization parameter through a gridsearch in the case of logistic classification and tweak parameters for the ID3 classifier such as the maximal depth etc.

7.3 Evaluating the model

Finally, we will analyze the model's performance using the metrics described in §5 and compare them against the benchmark.

References

[GLon] E. Gillich and V. Lohweg. Banknote authentication. *Conference Proceedings*, BVau(2010), https://www.researchgate.net/publication/266673146_Banknote_Authentication.